12

# ASPECTS OF GOODNESS-OF-FIT

Michael A. Stephens

*TECHNICAL REPORT No. 474*
*SEPTEMBER 30, 1993*

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

# ASPECTS OF GOODNESS-OF-FIT

Michael A. Stephens

*TECHNICAL REPORT No. 474*
*SEPTEMBER 30, 1993*

Prepared Under Contract
N00014-92-J-1264 (NR-042-267)
FOR THE OFFICE OF NAVAL RESEARCH

Professor Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305-4065

# Aspects of Goodness-of-Fit

MICHAEL A. STEPHENS

*Simon Fraser University, Burnaby, B. C., Canada V5A 1S6*

## Abstract

**In this article, two important methods of testing fit to a distribution are discussed and compared. They are the family of tests based on the empirical distribution function of a random sample, and the family based on plotting the order statistics against a suitable set of constants and examining the fit of a line through the plotted points. The two sets will be called EDF tests and Regression tests respectively.**

*Key Words:* Correlation tests; EDF tests; Probability plot; Regression tests; Tests for exponentiality; Tests for normality.

## 1 The goodness-of-fit problem

Suppose a random sample of $n$ values $x_1, x_2, x_3, \ldots, x_n$ is given, and it is desired to test that the sample comes from the distribution $F(x;\theta)$. The parameter $\theta$ represents a vector of parameters in the distribution; they may all be known, so that the tested distribution is completely specified — this situation will be called Case 0 — or some or all of the parameters may have to be estimated from the sample. Thus a test might be required of the hypothesis that the sample comes from a normal distribution with mean $\mu$ and variance $\sigma^2$, or that a sample comes from a Gamma distribution with scale parameter $\beta$ and shape parameter $m$. For the present, we assume the distribution $F(x;\theta)$ is continuous. We shall sometimes write the distribution as $F(x)$ for brevity.

# 2 EDF tests

The empirical distribution function of the sample is defined as follows:

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}, \\ i/n, & x_{(i)} \le x < x_{(i+1)},\ i = 1, 2, \ldots, n-1, \\ 1, & x \ge x_{(n)}. \end{cases}$$

The EDF thus represents, for any value of $x$, the fraction of the observations less than or equal to $x$; it clearly parallels $F(x)$, which gives the probability that an observation is less than $x$. In fact, by the Glivenko-Cantelli lemma, $|F_n(x) - F(x)| \to 0$ as $n \to \infty$. In 1933 Kolmogorov proposed a test based on the discrepancy $z_n(x) = F_n(x) - F(x)$, and Smirnov followed by proposing two related tests. The Kolmogorov-Smirnov tests, as they have come to be called, are defined as:

$$D_+ = \sup_x\{z_n(x)\};\ D_- = \sup_x\{-z_n(x)\}.$$

The statistic actually introduced by Kolmogorov was $D = \max(D_+, D_-)$. At about the same time, Cramér and von Mises were considering tests based on the integral of $z_n(x)$. The Cramér-von Mises family of statistics is

$$C = n \int_{-\infty}^{\infty} \{z_n(x)\}^2 \psi(x)\, dF(x),$$

where $\psi(x)$ is a weight function which can be used to vary the importance of different parts of the $x$-axis. Two commonly-used weight functions are $\psi(x) = 1$, giving the Cramér-von Mises statistic $W^2$, and $\psi(x) = \{F(x)[1 - F(x)]\}^{-1}$, giving the Anderson-Darling statistic $A^2$. In addition, $W^2$ can be modified to yield Watson's statistic $U^2$ given by

$$U^2 = n \int_{-\infty}^{\infty} \left\{F_n(x) - F(x) - \int_{-\infty}^{\infty} [F_n(x) - F(x)]\, dF(x)\right\}^2 dF(x).$$

## 2.1 Computing formulas

The definitions of these statistics look rather difficult to handle, but in fact very easy computing formulas exist. They are derived by means of the Probability Integral Transformation (PIT). This is the transformation

$$z = F(x; \theta).$$

It is well known that this transformation gives a variable $z$ which is uniformly distributed between 0 and 1, written $U(0,1)$. If the Kolmogorov-Smirnov and Cramér-von Mises statistics are now calculated from the EDF of the $z$-values, with $F(z) = z$, the uniform distribution, it may easily be shown that the values are the same as those calculated from the original $x$-diagram. The $z$-diagram then gives the computing formulas following, with $z_{(i)} = F(x_{(i)}, \theta)$:

$$\begin{aligned}
D^+ &= \max[i/n - z_{(i)}]; \\
D^- &= \max[z_{(i)} - (i-1)/n]; \\
D &= \max(D^+, D^-); \\
W^2 &= \frac{1}{12n} + \sum_i \left\{ z_{(i)} - \frac{2i-1}{2n} \right\}^2 ; \\
U^2 &= W^2 - n(\bar{z} - 0.5)^2 \text{ (where } \bar{z} = \textstyle\sum_i z_{(i)}/n); \\
A^2 &= -n - \frac{1}{n} \sum_i (2i-1)[\log z_{(i)} + \log(1 - z_{(n+1-i)})]. \qquad (1)
\end{aligned}$$

## 2.2 Estimated parameters

Suppose one or more components of $\theta$ are unknown, but are estimated by an efficient method from the sample values. These values are then inserted where necessary in the PIT above, and the statistics are calculated from the resulting $z$-values using the formulas (1). The unordered $z$-values are not now uniformly distributed; we describe them as superuniform, because they almost always give much smaller values for the statistics, implying that the $z$-values are more evenly spaced than a genuine uniform sample.

## 2.3 Another transformation to uniformity

It is well-known that if events are occurring randomly in time, say at times $t_1, t_2, \ldots, t_n$, (the clock is started at time zero), and if the values are transformed by $z_{(i)} = t_{(i)}/t_{(n)}$, the set of $n-1$ values $z_{(i)}, i = 1, 2, \ldots, n-1$, will be distributed $U(0,1)$. An interesting set of events which gives superuniform $z_{(i)}$ are the ends of reigns (deaths or abdications) of the Kings and Queens of England, starting with time zero as the accession of William I in 1066 — it is hard to explain this phenomenon, even though it is obvious that successive reigns have lengths which are correlated: see Pearson [1].

## 2.4 Distribution theory

When the continuous distribution tested is completely specified (this is called Case 0), so that the test of fit becomes a test that the $z$-values are uniformly distributed, percentage points of the EDF statistics are either known exactly, or can be approximated very accurately. Details and tables are given by Stephens [2]. Furthermore, it is possible to modify the statistics so that only the asymptotic points need be tabulated. To do this, a modified form $T^*$ of the EDF statistic $T$ is used which is an easily calculated function of $T$ and the sample size $n$. The resulting $T^*$ is then compared with the asymptotic points of the test statistic. Modified forms and tables are given in Biometrika Tables for Statisticians, Vol II, Table 54, and also in Stephens [2].

When parameters are estimated efficiently (that is, with asymptotic variances given by the inverse of the Fisher information matrix), and used in the PIT to give the $z$-values from which the statistics are calculated, asymptotic percentage points can be calculated for statistics of the Cramér-von Mises family. These include $W^2$, $U^2$, and $A^2$. The asymptotic points depend on the distribution being tested , but not on location or scale parameters in the distribution; however, they do depend on shape parameters such as occur in the Gamma or Weibull or von Mises distributions. The points for finite $n$ would be very difficult to calculate, and would have to be determined by Monte Carlo methods. Fortunately, for these statistics, the finite-$n$ points converge very quickly to the asymptotic points, so that the latter may be used for practical purposes — a test with very small sample size would in any case have very little power.

For Kolmogorov-Smirnov statistics the distribution theory is more difficult. Again, points will not depend on the true values of location or scale parameters, but even asymptotic points are very difficult to calculate. Such tables as exist have usually been found by Monte Carlo methods. In addition, points for finite $n$ do not converge rapidly to the asymptotic points for these statistics, so that it is necessary to give either the finite-$n$ points (obtained by Monte Carlo) or modified forms, as was done for Case 0.

For both families of statistics, extensive tables of points are given by Stephens [2] for testing for the normal, exponential, Gamma, Weibull, extreme-value, von Mises and Cauchy distributions, so that the tests are available for practical use.

## 2.5 Power

The power of a test statistic will of course depend on several factors, including the size (or $\alpha$-level) of the test, the sample size, and especially on the alternative to the tested distribution. Nevertheless, some general remarks can be made concerning the power of EDF statistics:

1. As two-sided omnibus tests (that is, tests against all alternatives, or at least a wide range of alternatives), the Cramér-von Mises family is more powerful in general than the Kolmogorov-Smirnov family. This might be expected, as the former "tests" the hypothesized distribution all along the range of values of $x$, while the latter looks for a marked discrepancy between the EDF and the hypothesized $F(x)$, possibly only around one point.

2. For Case 0, there is a difference in power between the statistics, according to whether the alternative distribution is mostly a change in the location of the distribution, or

a change in the scale. $W^2$ and $A^2$ will detect a change in location, and $U^2$ a change in scale. The Kolmogorov statistic $D$ also detects a change in location.

3. If there is a change in location, ***and the direction is known***, the statistics $D^+$ or $D^-$ can be very powerful; however, if the wrong statistic is used, the power can easily be less than the $\alpha$-level — that is, the test is biased. $D^+$ detects the situation where the true location is less than that tested, and $D^-$ detects the opposite situation.

4. When parameters are estimated, the differences between the powers for these various types of alternative tend to fade, although the Cramér-von Mises family will still be better overall than the Kolmogorov-Smirnov tests.

5. On the whole, the recommended test statistic is the Anderson-Darling $A^2$; it is particularly effective in detecting outliers, that is, observations which are further into the tails than expected, and this is often the situation which the tester most wishes to detect.

Further details on all these statistics are given by Stephens [2]; a discussion of their use, and comparisons with other statistics, for the "observations random in time" situation described briefly above, is in Stephens [3].

# 3 Regression tests

## 3.1 Introduction

For the second part of this paper, we describe another group of tests, to be called regression tests. They are based on a well-established and popular technique for testing fit to selected distributions, the ***probability plot***. In regression tests, the order statistics $x_{(i)}$ of a sample are plotted on the vertical axis of a graph, against $t_i$, a set of constants which depend only on $i$, along the horizontal axis. (In the probability plot, the axes were reversed, but for convenience in introducing test statistics we keep them as above). The constants $t_i$ are chosen so that the relationship between the $x_{(i)}$ and $t_i$ is approximately a straight line. Historically, the linear relationship was often judged by eye, but more recently, test statistics have been developed, based on the parameters associated with the straight-line fit, when this is done by ordinary or generalised least squares.

Regression tests arise naturally when unknown parameters in the tested distribution $F(x;\theta)$ are location and scale parameters. Suppose $F(x;\theta)$ is $F_0(w)$, where $F_0(w)$ is a completely specified distribution and $w = (x-\alpha)/\beta$; then $\theta = (\alpha, \beta)$ with $\alpha$ a location parameter and $\beta$ a scale parameter. A sample with order statistics $x_{(i)}$ can be derived from a set of values $w$ from $F_0(w)$ with order statistics $w_{(i)}$, by the relationship

$$x_{(i)} = \alpha + \beta w_{(i)}, \qquad i = 1, \ldots, n. \tag{2}$$

An obvious example is the test for normality, where the density of $w$ is given by $f(w) = (2\pi)^{-1/2}\exp(-w^2/2)$. Let $\Phi(w) = \int_{-\infty}^{w} f(t)\,dt$; then $F_0(w) = \Phi(w)$ and $F(x;\theta) = \Phi(w)$ with $w = (x-\mu)/\sigma$.

In the more general case, let $m_i = E(w_{(i)})$; then, from (2) we have

$$E(x_{(i)}) = \alpha + \beta m_i \tag{3}$$

and a plot of $x_{(i)}$ against $m_i$ should be approximately a straight line with intercept $\alpha$ on the vertical axis and slope $\beta$. The values $m_i$ are the most natural values to plot along the horizontal axis, but for most distributions they are difficult to calculate. Various authors have therefore proposed alternatives $t_i$ which are convenient functions of $i$; then (3) can be replaced by the model

$$x_{(i)} = \alpha + \beta t_i + \epsilon_i \tag{4}$$

where $\epsilon_i$ is an "error" which only for $t_i = m_i$ will have mean zero.

It is then important to find a good method of testing how well the data fits the line (3) or (4). One way is simply to measure the correlation coefficient $r(x,t)$ between the paired sets $x_{(i)}$ and $t_i$. A second method is to estimate $\beta$ using generalised least squares, and to compare this estimate with the estimate of scale given by the sample variance. We now examine these two procedures.

## 3.2 The correlation coefficient as test statistic

In discussing the correlation coefficient $r(x,t)$, we extend the usual meaning of correlation, and also that of variance and covariance, to apply to constants as well as random variables. Thus let $x$ refer to the vector $x_{(1)},\ldots,x_{(n)}$, and $t$ to the vector $t_1,\ldots,t_n$; let $\bar{x} = \sum x_{(i)}/n$ and $\bar{t} = \sum t_i/n$, and define the sums

$$\begin{aligned} S(x,t) &= \sum(x_{(i)}-\bar{x})(t_i-\bar{t}) = \sum x_{(i)}t_i - n\bar{x}\bar{t} \\ S(x,x) &= \sum(x_{(i)}-\bar{x})^2 = \sum(x_i-\bar{x})^2 \\ S(t,t) &= \sum(t_i-\bar{t})^2. \end{aligned}$$

$S(x,x)$ will often be called $S^2$.

The correlation coefficient between $x$ and $t$ is

$$r(x,t) = \frac{S(x,t)}{[S(x,x)S(t,t)]^{1/2}}. \tag{5}$$

Statistics $r(x,m)$ or $r^2(x,m)$ are natural statistics for testing the fit of $x$ to the model (3), since if a "perfect" sample is given, that is, a sample whose ordered values fall exactly at their expected values, $r(x,m)$ will be 1; more generally, the value of $r(x,m)$ can be interpreted as a measure of how closely the sample resembles a perfect sample. Tests based on $r(x,m)$, or equivalently on $r^2(x,m)$, will be one-tailed, with rejection of $H_0$ occurring only for low values of $r$.

However, as $n \to \infty$, $r^2(x,m) \to 0$ on $H_0$. A statistic which does have an asymptotic distribution is

$$Z(x,m) = n\{1 - r^2(x,m)\}. \tag{6}$$

Then $Z(x,m)$ is an equivalent statistic to $r^2$, based on the sum of squares of the residuals after the line (3) has been fitted. In common with many other goodness-of-fit statistics, for example chi-square and the EDF statistics, $Z(x,m)$ has the property that the larger it is, the worse the fit. Sarkadi [4] showed consistency of the test based on $r(x,m)$ for normality, and Gerlach [5] has shown consistency for correlation tests based on $r(x,m)$, or equivalently $Z(x,m)$, for a wide class of distributions including all the usual continuous distributions. This is to be expected, since, for large $n$, we can expect our sample to become perfect in the sense above. We can expect the consistency property to extend to $r(x,t)$ provided that $t$ approaches $m$ sufficiently rapidly for large samples.

## 3.3 The correlation test for the normal distribution

For the normal distribution $N(\mu,\sigma^2)$, $f(w) = (2\pi)^{-1/2}\exp(-w^2/2)$, with $w = (x-\mu)/\sigma$; thus $\alpha = \mu$ and $\beta = \sigma$, and the $m_i$ are the expected values of standard normal order statistics. Equation (3) becomes

$$E(x_{(i)}) = \mu + \sigma m_i. \tag{7}$$

For the normal distribution $\bar{m} = 0$, and $r^2(x,m)$ can conveniently be written in vector notation. Let $x$ be the vector $(x_{(1)},\ldots,x_{(n)})$, and let $m$ be the vector $(m_1,\ldots,m_n)$; let primes, eg. $x'$ and $m'$, denote transposes of vectors or matrices.

Then

$$r^2(x,m) = \frac{(x'm)^2}{(m'm)S^2}. \tag{8}$$

The values of $m_i$ required for the calculation of $r^2(x,m)$ have been well tabulated, and good computer programs are also available.

This statistic will later on be seen to be identical to $W'$, the Shapiro-Francia statistic, so that, for testing normality, we shall refer to $r^2(x,m)$ also as $W'$. Tables for $W'$ have been given by Shapiro and Francia [6].

In practice, it is easier to interpolate in tables of $Z(x,m)$ rather than $r^2(x,m)$, and Stephens [7] has produced tables for $Z(x,m)$ for both complete and censored samples. The null hypothesis that the sample comes from a normal distribution is rejected for large values of $Z(x,m)$.

De Wet and Venter [8] have proposed the use of the statistic $r(x, H)$, where $H_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$. Use of $H_i$ makes distribution theory easier, and de Wet and Venter have given the asymptotic null distribution of $Z(x, H)$. The $H_i$ must be found numerically, using one of the excellent approximations available for $\Phi^{-1}(\cdot)$. The values of $H_i$ and $m_i$ are close in the middle of the sample, but are wider apart at the extremes. However, in Leslie *et. al.* [9] it is shown that $Z(x, H)$ and $Z(x, m)$ have the same null asymptotic distributions.

## 3.4 The Shapiro-Wilk procedure

We next turn to the second method of testing mentioned above, in which the parameters $\alpha$ and $\beta$ in the model $x_{(i)} = \alpha + \beta m_i$ are estimated by generalised least squares. Using our previous notation, let $w_{(i)}$ be the order statistics from $F(w)$ with $\alpha = 0$ and $\beta = 1$; let $m_i = E(w_{(i)})$ as before, and let $E(w_{(i)} - m_i)(w_{(j)} - m_j) = V_{ij}$, the covariance of $w_{(i)}$ and $w_{(j)}$. Then let $x$ be the column vector with components $x_{(1)}, \ldots, x_{(n)}$, let $m$ be a column vector with components $m_1, \ldots, m_n$, and let 1 be a column vector with each component equal to 1. Let $V$ be the matrix with elements $V_{ij}$. The generalised least squares estimates of $\alpha$ and $\beta$ are then

$$\hat{\alpha} = -m'Gx \text{ and } \hat{\beta} = 1'Gx, \tag{9}$$

where

$$G = \frac{V^{-1}(1m' - m1')V^{-1}}{(1'V^{-1}1)(m'V^{-1}m) - (1'V^{-1}m)^2}. \tag{10}$$

For some distributions, for example the normal and exponential, these equations simplify considerably.

A method of testing fit has been proposed by Shapiro and Wilk [10, 11] for testing normality and exponentiality. The procedure used is basically to compare the estimate of $\beta^2$ given by equation (9) with the estimate of $\beta^2$ given by the sample variance; the ratio of these estimates, multiplied by a constant, is taken as the test statistic. In the case of tests for normality, slight modifications of the first estimate of $\beta^2$ have also been suggested, since the estimate is complicated to calculate.

For the Shapiro-Wilk test for normality, $\alpha$ and $\beta$ in (3) are $\mu$ and $\sigma$ respectively; the estimates of these parameters given by (9) then become

$$\hat{\mu} = \bar{x} \text{ and } \hat{\sigma} = \frac{m'V^{-1}x}{m'V^{-1}m}.$$

The test statistic proposed by Shapiro and Wilk [10] is

$$W = \frac{\hat{\sigma}^2 R^4}{S^2 C^2} \tag{11}$$

where $S^2 = \sum(x_{(i)} - \bar{x})^2 = \sum(x_i - \bar{x})^2$, $R^2 = m'V^{-1}m$, and $C^2 = m'V^{-1}V^{-1}m$. The factors $R^4$ and $C^2$ ensure that W always takes values between 0 and 1.

Suppose the vector $a$ is defined by $a = V^{-1}m/C$; then

$$W = \frac{(a'x)^2}{S^2} = \frac{(\sum_i a_i x_{(i)})^2}{S^2}.$$

In order to calculate $W$, the vector $a$ is needed, and this in turn requires values of $m$ and $V^{-1}$, derived from $V$. For values of $n$ between 21 and 50, Shapiro and Wilk used approximations for the components $a_i$ of $a$, and gave a table of values of $a_i$ for sample sizes from $n = 3$ to 50. They also gave Monte Carlo points with which to make the test. The test is one-tailed: small values of $W$ are significant.

A test similar to $W$, but for use with $n \geq 50$, was later suggested by Shapiro and Francia [6]. This is based on the observation of Gupta [12], who noted that the estimate $\hat{\sigma}$ is almost the same if $V^{-1}$ is ignored in equation (10); the test statistic then given by Shapiro and Francia is

$$W' = \frac{(m'x)^2}{(m'm)^2 S^2}.$$

As has already been observed, this is equivalent to the sample correlation statistic $r^2(x,m)$.

### 3.5 Asymptotic equivalence of the Shapiro-Wilk and correlation statistics

Thus we have the remarkable result that the Shapiro-Wilk statistic *for testing normality* approaches the correlation coefficient $r^2(x,m)$. It is interesting to ask why this is so: why $V^{-1}$ can be "ignored" when calculating $W$. Stephens [13] has shown heuristically that, for large $n$, $m$ becomes an eigenvector of $V$, and $Vm \to \frac{1}{2}m$; then $V^{-1}m \to 2m$, $m'V^{-1}x \to 2m'x$, and $m'V^{-1}m \to 2m'm$. Hence $W \to W'$ because the factor 2 cancels in the numerator and denominator of $W$. The above results were proved rigorously by Leslie [14]. Stephens [13] also gives other asymptotic eigenvalues and eigenvectors of $V$.

## 4 Power comparisons

Shapiro and Wilk [10] gave power results for $W$, based on Monte Carlo studies. Unfortunately, the comparisons with EDF statistics were inaccurate — the EDF statistics were compared with Case 0 tables, and not the Case 3 tables to be used when the parameters $\mu$ and $\sigma$ are estimated by $\bar{x}$ and $s$. Stephens [15] later gave comparisons based on the correct tables. These show that $W$ is barely superior overall to EDF statistics, and especially only slightly superior to the Anderson-Darling $A^2$. Both statistics tend to have higher power than older statistics such as $b_1$ and $b_2$, the coefficients of skewness and kurtosis.

# 5 Two questions

The above results show that $W$, equivalent to the correlation coefficient $r^2(x,m)$, appears to give overall the most powerful omnibus test *for normality*. Two questions can then be asked:

(a) Will the Shapiro-Wilk procedure give good tests for other distributions?

(b) Will the correlation coefficient be successful for testing other distributions?

With reference to the first question, we first observe that for other distributions tested, the Shapiro-Wilk procedure will not necessarily lead to a statistic which is asymptotically equivalent to the correlation coefficient.

For the exponential distribution, where $F(x;\theta) = 1 - \exp\{-(x-\alpha)/\beta\}$, provided $x \geq \alpha$, (thus $F(w) = 1 - \exp(-w)$, and $\theta = (\alpha, \beta)$), the estimates in (3) become

$$\hat{\alpha} = x_{(1)} \text{ and } \hat{\beta} = \frac{n(\bar{x} - x_{(1)})}{(n-1)}.$$

The ratio $\hat{\beta}^2/S^2$, omitting some factors involving $n$, leads to the statistic

$$W_E = \frac{n(\bar{x} - x_{(1)})^2}{(n-1)S^2}.$$

Although this statistic has been proposed, and points given, for testing exponentiality (Shapiro and Wilk [11], Currie [16]), it has not proved powerful (Stephens [3]). Furthermore, it does not provide a *consistent* test, which means that there will be some distributions, not exponential, which would not be detected with power approaching 1, when the test for exponentiality is applied to large samples. Sarkadi [4] first pointed this out, by observing that $W_E$ is equivalent, for large samples, to the coefficient of variation (CV) of the sample. To fix ideas, suppose $\alpha$ is known to be zero (this is frequently the case when the exponential distribution is used, although the discussion which follows is easily adapted to the case where $\alpha$ is not zero). Then, for large $n$, $x_{(1)} \to \alpha = 0$, and $W_E \to \bar{x}^2/S^2$. The coefficient of variation is $S^2/\bar{x}^2$, so that $W_E \to 1/\text{CV}$, and for large $n$, this is 1. However, many distributions have CV = 1, and a very large sample from one of these will have a $W_E$ also approaching 1. The power of $W_E$ will then approach a constant (less than 1) depending on the variance of $W_E$.

Spinelli and Stephens [17] have given power studies with samples taken from some other distributions with CV = 1, where the power of $W_E$ is seen to diminish as the sample size $n$ increases. Lockhart and Stephens [18] have explored the question of non-consistency further, and have shown that only for a very limited family of distributions, including the normal, does the Shapiro-Wilk procedure give a consistent test. Thus this technique of basing a test on the ratio of the regression estimate of scale to that given by the sample standard deviation cannot be recommended except for the normal case.

We now turn to the second question above. Since the correlation coefficient is powerful for testing normality, will it be equally successful for tests on other distributions? First, it should be emphasised that the most appropriate correlation is that between $x$ and $m$: in the normal case, $H_i = \Phi^{-1}\{i/(n+1)\}$ was "sufficiently close" to $m$ that the correlations $r(x, H)$ and $r(x, m)$ were approximately equal for large samples, and so had the same power. This is *not so* for other distributions. For example, for the exponential distribution, $m_i = \sum_{j=1}^{i}(n+1-j)^{-1}$, and $H_i = -\log\{1 - i/(n+1)\}$, and these are not close enough in the tails to give $r(x, H)$ as much power as $r(x, m)$. We can expect this result to be true also for other long-tailed distributions, and the question of when $r(x, H)$ can replace $r(x, m)$ for those distributions for which $m$ is hard to calculate is itself an interesting research topic. See, for example, McLaren and Lockhart [19].

We therefore confine further discussion to the properties of $r(x, m)$. We have seen that, in contrast to $W$, $r(x, m)$ always gives a consistent test. However, McLaren and Lockhart [19] show that the asymptotic relative efficiency of correlation tests can be zero compared with EDF tests.

Stephens [20] adapted $W_E$ to test exponentiality in the case where $\alpha$ is known; this can be compared with most power studies on other statistics which usually assume $\alpha = 0$. Stephens Ste86c gives some tables for comparison, and these demonstrate that the $W$ statistics are in general less powerful than EDF statistics.

## 6 Censored data

One attraction of correlation statistics is the fact that the correlation coefficient is well-known to most applied statisticians, and the formula is very easy to calculate. This is true also for censored observations of types I or II, where missing observations are all at one end of the sample, often in the right-hand tail where higher values occur. Because of this appeal, Stephens [7], as was stated earlier, gives many tables of $Z(x, m) = n\{1 - r^2(x, m)\}$, or of the corresponding $Z(x, H)$, for use with right-censored data and for testing the exponential, Weibull and other distributions. EDF statistics have also been adapted for censored data, and formulas and tables for these statistics are given by Stephens [2]. For censored data, as for full samples, the statistics $Z(x, m)$ and $Z(x, H)$ may not be as powerful in general as EDF statistics; much depends on the influence of the tail observations which are lost by censoring. Finally, randomly censored data poses a unique problem in testing fit. The Kaplan-Meier estimate of $F(x)$ can be used for EDF statistics, and $r^2(x, m)$ can still be calculated if it is known *which* ordered observations have been lost, but in either case tables are difficult to provide. More work is needed on this topic.

# 7 Tests for discrete distributions

Until now, tests have been discussed only for continuous distributions. The correlation coefficient and the Shapiro-Wilk procedure do not adapt readily to discrete distributions, but EDF tests can be adapted. The technique is based on measures of discrepancy between the cumulative histograms of observed values and expected values (Pearson's $\chi^2$ measures the discrepancy within each cell, and does not sum the observeds and expecteds). Pettitt and Stephens [21] gave some distribution theory for the Kolmogorov-Smirnov statistic for testing uniformity, and Freedman [22] discussed the Watson $U^2$ statistic, one of the Cramér-von Mises family, for the discrete uniform test. Recently, Lockhart and Stephens [23] have extended the test to include $W^2$ and $A^2$, and Spinelli and Stephens [24] develop a test for the Poisson distribution using Cramér-von Mises statistics. Power studies show these tests to be quite effective. In particular, for the test for normality, the EDF statistics will be more powerful than Pearson's $\chi^2$ when the alternative is a trend in the cell probabilities — for example, to test that the probability of a defective item produced in a factory is the same each week, against the alternative that it decreases with time.

# 8 Summary and final remarks

In this paper we have reviewed two important methods of testing fit — EDF statistics and regression methods based on the probability plot. Tests based on the EDF and those based on the correlation coefficient $r(x,m)$ are *consistent*, whereas those derived from use of the Shapiro-Wilk procedure are *not consistent* in general. The exception is the test for normality. For large samples, the correlation coefficient, however, can have low efficiency compared with EDF tests. For smaller samples, and for censored data, the situation is less clear, and more work is needed. Of course, other techniques for testing fit exist, based on Pearson's $\chi^2$, on spacings, or on the empirical characteristic function. In general, for tests for continuous distributions, Pearson's $\chi^2$ has low power compared with EDF statistics, due to the loss of information resulting from the grouping required. EDF statistics compare well with the other methods also, and, for overall testing against omnibus alternatives, these statistics are recommended. For specified *limited* alternatives, clearly other tests (for example the Likelihood Ratio test) can have good properties. Stephens [2, 3, 7] discusses these issues, but much more research can be done, both on mathematical aspects of the statistics, and on practical comparisons of tests.

# References

[1] E. S. Pearson. Comparison of tests for randomness of points on a line. *Biometrika* **50** 315–325 (1963).

[2] M. A. Stephens. Tests based on EDF statistics. Chapter 4 in *Goodness-of-fit techniques* (R.B. d'Agostino and M.A. Stephens, eds.). New York: Marcel Dekker (1986).

[3] M. A. Stephens. Tests for the exponential distribution. Chapter 10 in *Goodness-of-fit techniques* (R.B. d'Agostino and M.A. Stephens, eds.). New York: Marcel Dekker (1986).

[4] K. Sarkadi. The consistency of the Shapiro-Francia test. *Biometrika* **62** 445–450 (1975).

[5] B. Gerlach. A consistent correlation-type goodness-of-fit test; with application to the two-parameter Weibull distribution. *Math. Operationsforsch. Statist. Ser. Statist.* **10** 427–452 (1979).

[6] S. S. Shapiro and R. S. Francia. Approximate analysis of variance test for normality. *J. Amer. Statist. Assoc.* **67** 215–216 (1972).

[7] M. A. Stephens. Tests based on regression and correlation. Chapter 5 in *Goodness-of-fit techniques* (R.B. d'Agostino and M.A. Stephens, eds.). New York: Marcel Dekker (1986).

[8] T. De Wet and J. H. Venter. Asymptotic distribution of certain test criteria of normality. *South African Statist. J.* **6** 135–149 (1972).

[9] J. Leslie, S. Fotopoulos and M. A. Stephens. Asymptotic distribution of the Shapiro-Wilk $W$ for testing normality. *Annals of Statistics* 14 1497–1506 (1986).

[10] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika* **52** 591–611 (1965).

[11] S. S. Shapiro and M. B. Wilk. An analysis of variance test for the exponential distribution (complete samples). *Technometrics* 14 355–370 (1972).

[12] A. K. Gupta. Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika* **39** 266–273 (1952)

[13] M. A. Stephens. Asymptotic properties for covariance matrices of order statistics. *Biometrika* **62** 23–28 (1975).

[14] J. Leslie. Asymptotic properties and a new approximation for both the covariance matrix of normal order statistics and its inverse. *Colloq. Math. Soc. Janos Bolyai on Goodness of Fit* **45** (1984).

[15] M. A. Stephens. EDF statistics for goodness-of-fit and some comparisons. *J. Amer. Statist. Assoc.* **69** 730–737 (1974).

[16] I. D. Currie, The upper tail of the distribution of $W$-exponential. *Scand. J. Statist.* 7 147–149 (1980).

[17] J. J. Spinelli and M. A. Stephens. Tests for exponentiality when origin and scale parameters are unknown. *Technometrics* 29 471–476 (1987).

[18] R. A. Lockhart and M. A. Stephens. The non-consistency of the Shapiro-Wilk procedure. Research report, Dept. of Mathematics and Statistics, Simon Fraser University (1992).

[19] G. D. McLaren and R. A. Lockhart. On the asymptotic efficiency of certain tests of fit. *Canad. J. Statist.* 15 159–167 (1987).

[20] M. A. Stephens. On the $W$ test for exponentiality with origin known. *Technometrics* 20 33–35 (1978).

[21] A. N. Pettitt and M. A. Stephens. The Kolmogorov-Smirnov goodness-of-fit statistic with dicrete and grouped data. *Technometrics* 19 205–210 (1977).

[22] L. Freedman. Watson's $U^2$ statistic for discrete distributions. *Biometrika* 68 708–711 (1981).

[23] R. A. Lockhart and M. A. Stephens. Cramér-von Mises statistics for discrete distributions. Research report, Dept. of Mathematics and Statistics, Simon Fraser University (1992).

[24] J. J. Spinelli and M. A. Stephens. EDF tests for the Poisson distribution. Research report, Dept. of Mathematics and Statistics, Simon Fraser University (1992).

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| | | |
|---|---|---|
| **REPORT DOCUMENTATION PAGE** | | READ INSTRUCTIONS<br>BEFORE COMPLETING FORM |
| 1. REPORT NUMBER<br>474 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>ASPECTS OF GOODNESS OF FIT | | 5. TYPE OF REPORT & PERIOD COVERED<br>Technical |
| | | 6. PERFORMING ORG. REPORT NUMBER |
| 7. AUTHOR(s)<br>Michael A. Stephens | | 8. CONTRACT OR GRANT NUMBER(s)<br>N0025-92-J-1264 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Department of Statistics<br>Stanford University<br>Stanford, CA 94305-4065 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>NR-042-267 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Office of Naval Research<br>Statistics & Probability Program<br>Code 111 | | 12. REPORT DATE<br>September 30, 1993 |
| | | 13. NUMBER OF PAGES<br>16 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified |
| | | 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

THE VIEW, OPINIONS, AND/OR FINDINGS CONTAINED IN THIS REPORT ARE THOSE OF THE AUTHOR(S) AND SHOULD NOT BE CONSTRUED AS AN OFFICIAL DEPARTMENT OF THE ARMY POSITION, POLICY, OR DECISION, UNLESS SO DESIGNATED BY OTHER DOCUMENTATION.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

*Key Words:* Correlation tests; EDF tests; Probability plot; Regression tests; Tests for exponentiality; Tests for normality.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

In this article, two important methods of testing fit to a distribution are discussed and compared. They are the family of tests based on the empirical distribution function of a random sample, and the family based on plotting the order statistics against a suitable set of constants and examining the fit of a line through the plotted points. The two sets will be called EDF tests and Regression tests respectively.